Text Mining the Enron Corpus

Mihir Kelkar
Advised by : Dr. Charles Nicholas

Overview

Authorship Attribution

Overview

Overview

Authorship Attribution

Using Machine Learning to Find Financial Criminals

Overview

Authorship Attribution

</> Using Machine Learning to Find Financial Criminals

Analyzing Sentiment Trends within the communications

Brief History

Enron Corporation

Brief History

Enron Corporation

Northern Natural Gas Company

1930

Brief History

Enron Corporation

Brief History

Enron Corporation

Internorth Inc.

1970

Brief History

Enron Corporation

Brief History

Enron Corporation

Houston Natural Gas

Internorth Inc.

1985

Brief History

Enron Corporation

1985

Brief History

Enron Corporation

1985

Brief History

Enron Corporation

Brief History

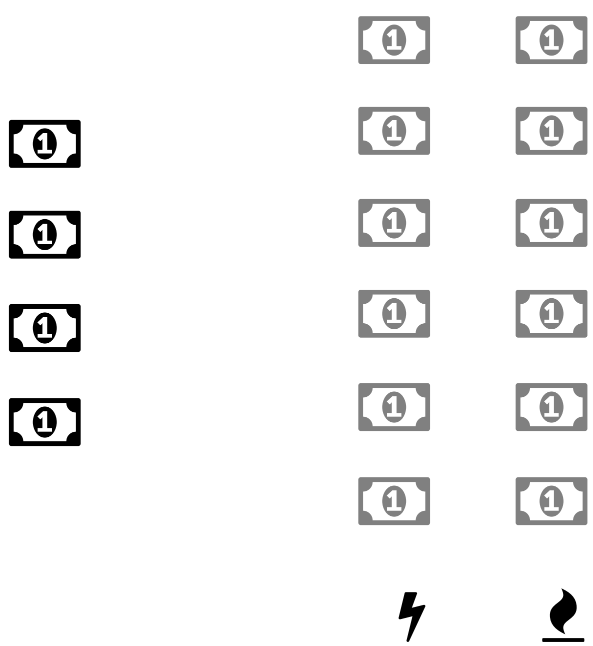
Enron Corporation

Brief History

Enron Corporation

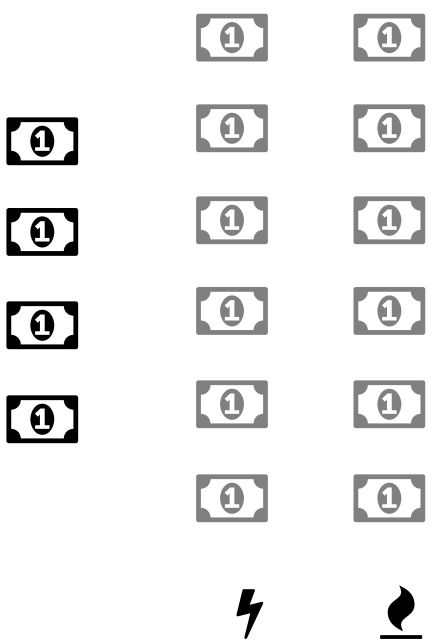
Brief History

Enron Corporation



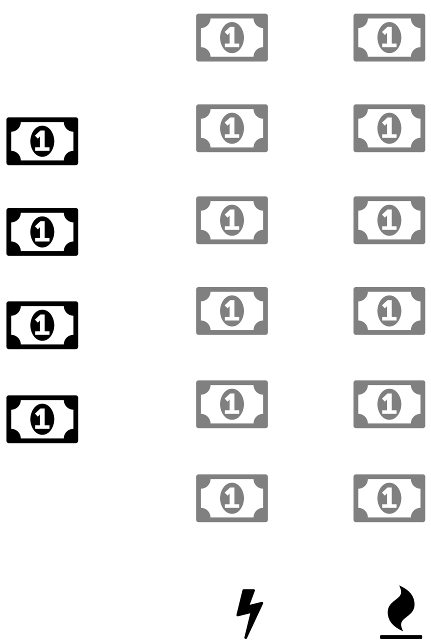
Brief History

Enron Corporation



Brief History

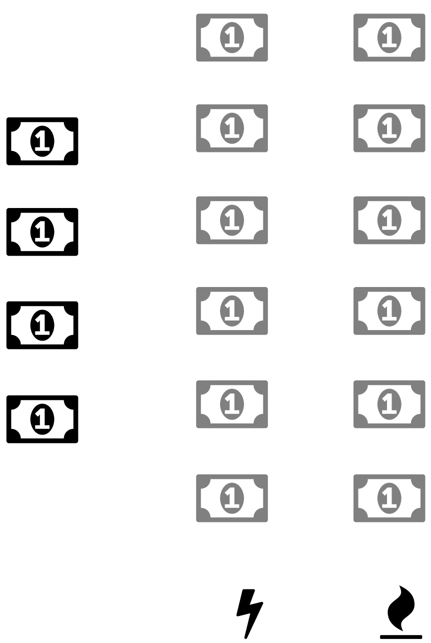
Enron Corporation

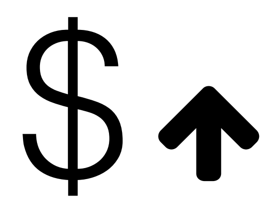


$

Brief History

Enron Corporation





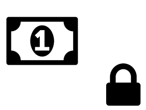
Brief History

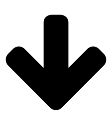
Enron Corporation

$

Brief History

Enron Corporation



$ 

Authorship Attribution

Half a Million Emails

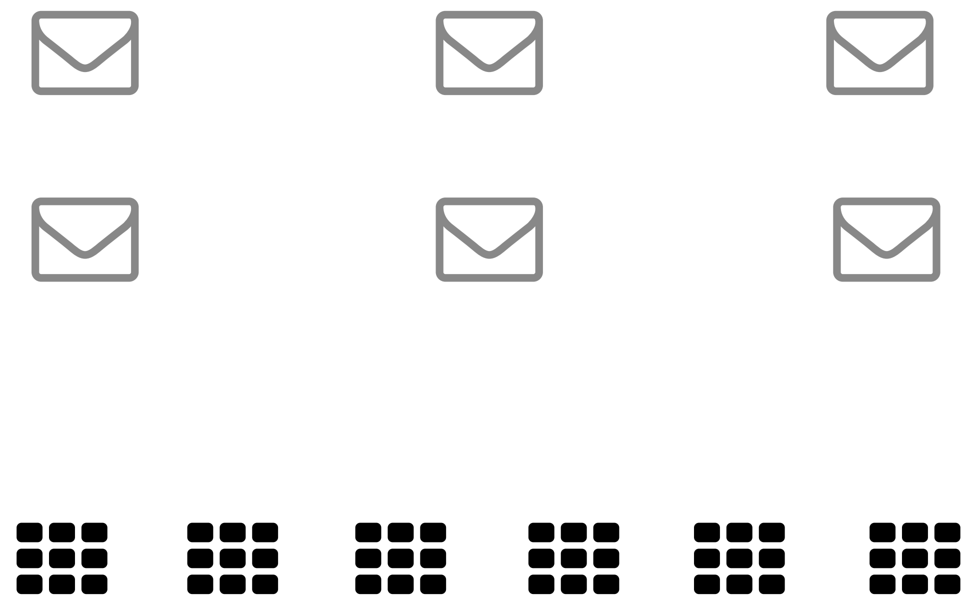
All 151 Top Executives of Enron

Over 4 GB after pre-processing

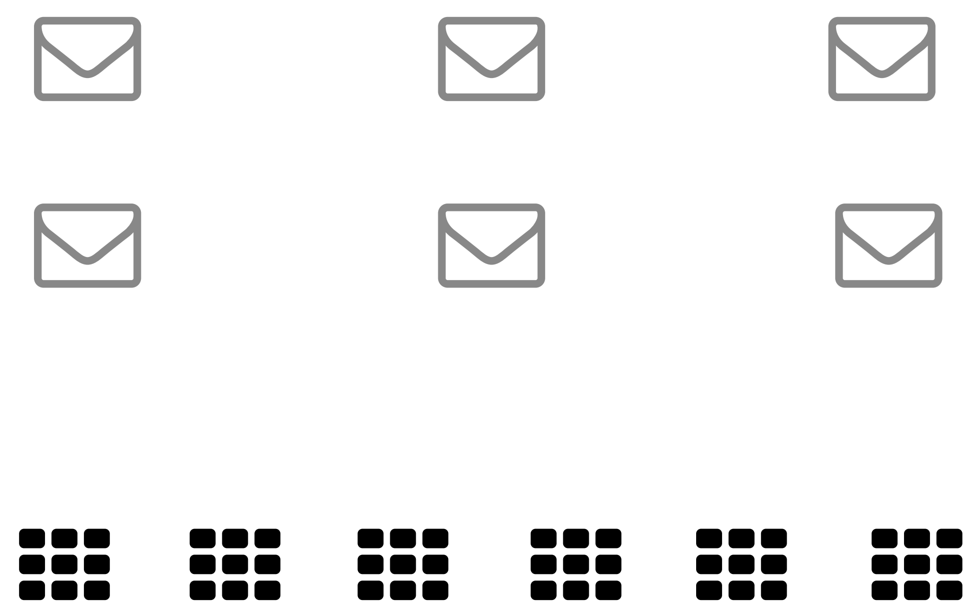
Primary Aim

Quantifying Emails from a User

Quantifying Emails from a User



Quantifying Emails from a User



{ } { } { } { }

Block Vectors

Quantifying Emails from a User

Block Vector

Quantifying Emails from a User

Block Vector

Function Words

"Enron is one of the biggest corporations in the world"

Quantifying Emails from a User

Block Vector

Quantifying Emails from a User

Bigrams

Block Vector

(Enron, is) (Enron, one)
(one, the) (Enron, World)

Quantifying Emails from a User

Block Vector

Quantifying Emails from a User

Block Vector

Trigrams*

(Enron, is, the) (Enron, one, of)
(one, world, the) (Enron, of, World)

*And Combinations of all three

Quantifying Emails from a User

Block Vector

Quantifying Emails from a User

{ } + { } + { } + { } == [ ]

Average User
Vector

Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Naive Method

Euclidian Distance

Manhattan Distance

Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Naive Method

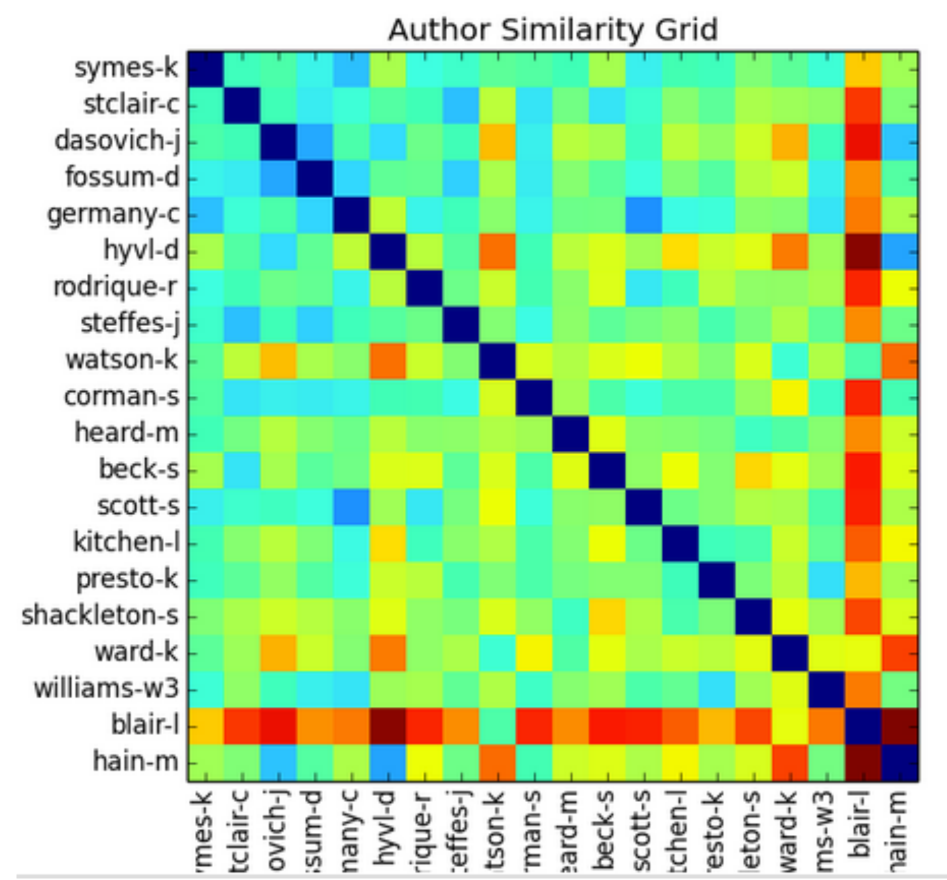
Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Naive Method

Bluer tones indicate more similarity



Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Better Method

Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Better Method

Perform SVD
Reduce dimensions

Use Cosine Similarity

Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

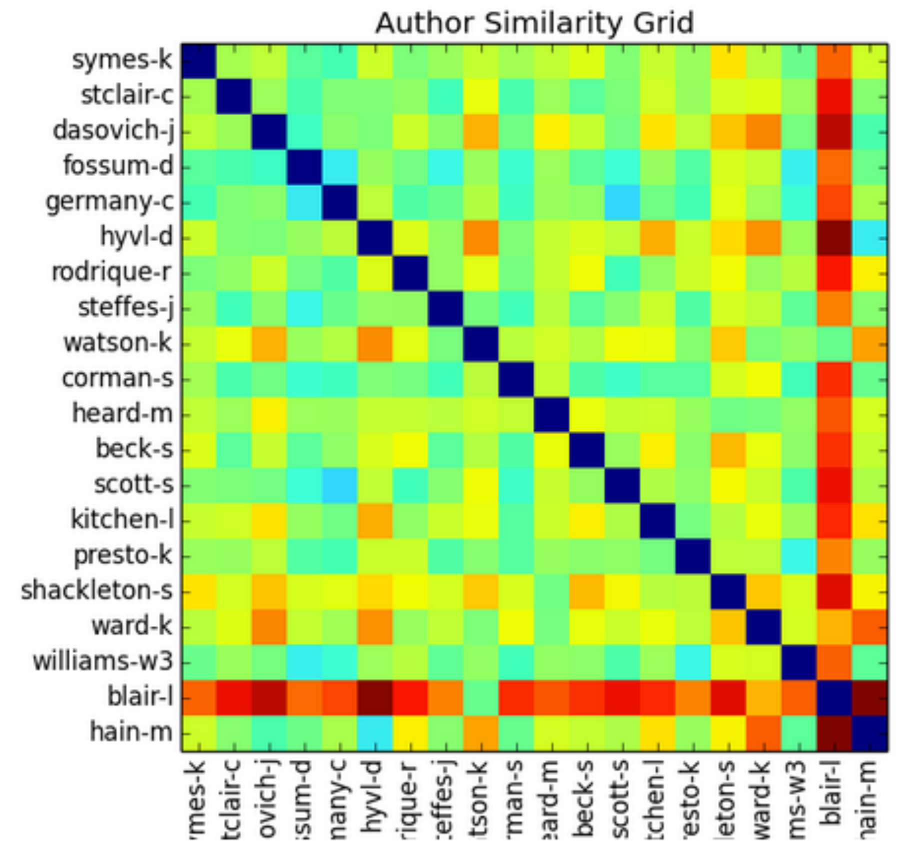
Better Method

Finding Similarities in Authors

We analyzed messages for the top 20 authors to find similarities in writing styles.

Each author was represented by their average user vectors

Better Method



Visualizing User Clusters

Less Divergence = Same Author

More Divergence = Multiple Authors
probably

Overlapping clusters = Authorship Anomaly

Visualizing User Clusters

TL;DR

Tightly bound clusters which are further away from other clusters indicate a single author

Visualizing User Clusters

$$\frac{\sum_{1}^{K} d(User\_Block_i - User\_Centroid_A)}{\sum_{1}^{M} d(User\_Centroid_A - User\_Centroid_j)}$$

* Cluster A has K word blocks

*Total of M clusters

Visualizing User Clusters

Visualizing User Clusters

Smaller the value,the more likelihood that
the cluster had a single author

Visualizing User Clusters

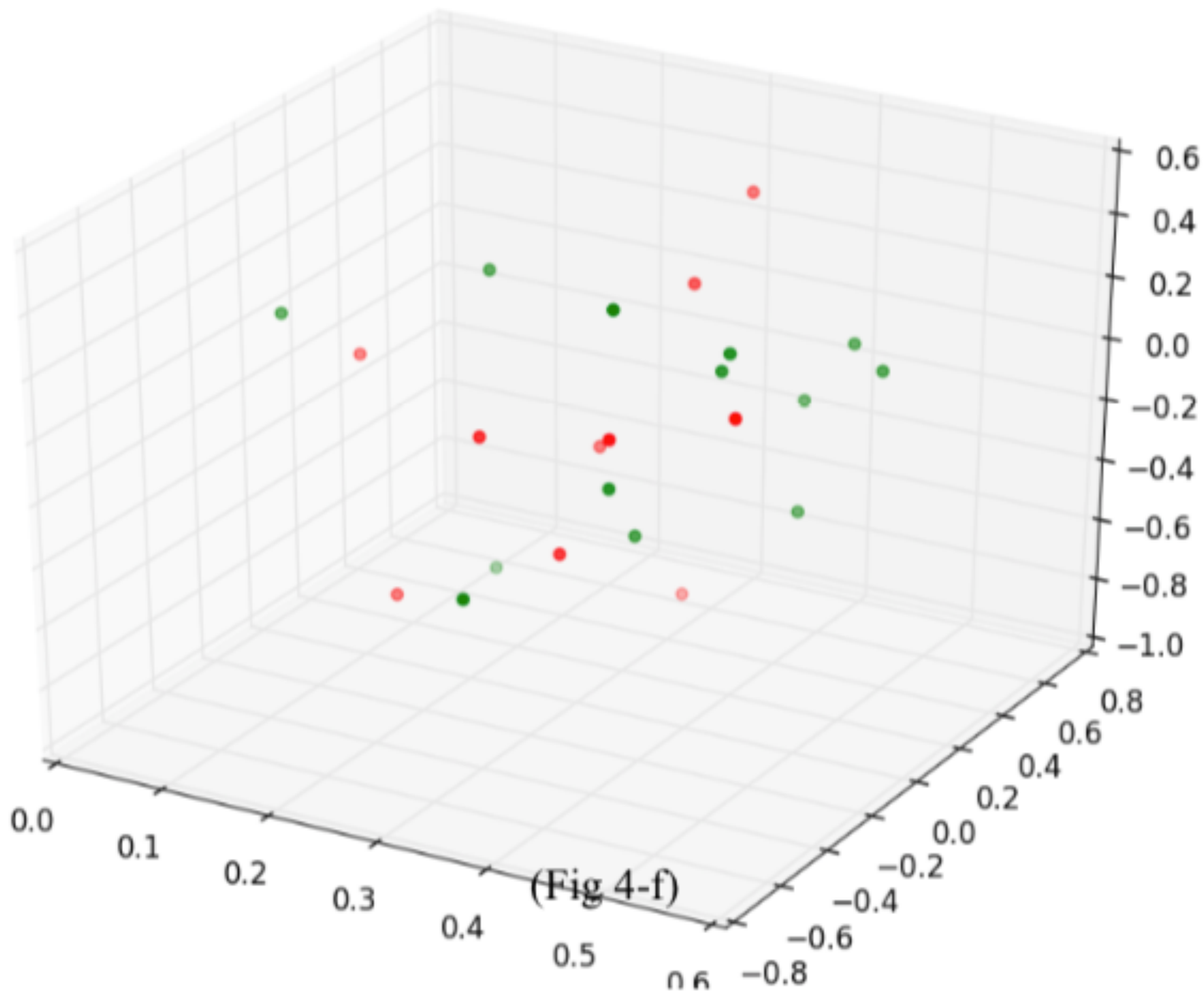
We visualized all *significant* authors in the corpus

Two Clusters stood out

Cluster for Kaminski, Vince

Cluster for Mann, Kay

Visualizing User Clusters



Word Block Size : 2000 words

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Dealt with FERC regulations

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Dealt with FERC regulations

Replied to people looking for jobs

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Dealt with FERC regulations

Replied to people looking for jobs

Frequently referenced Washington DC

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Dealt with FERC regulations

Replied to people looking for jobs

Frequently referenced Washington DC

Frequently referenced Universities

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Dealt with FERC regulations

Replied to people looking for jobs

Frequently referenced Washington DC

Frequently referenced Universities

Both Mr.Kaminski and Ms.Mann had raised objections about Enron’s dubious financial practices. However, we did not see any conclusive evidence that either of them worked together on exposing this.

Visualizing User Clusters

We believe that the similarity in both clusters exists because

Dealt with FERC regulations

Replied to people looking for jobs

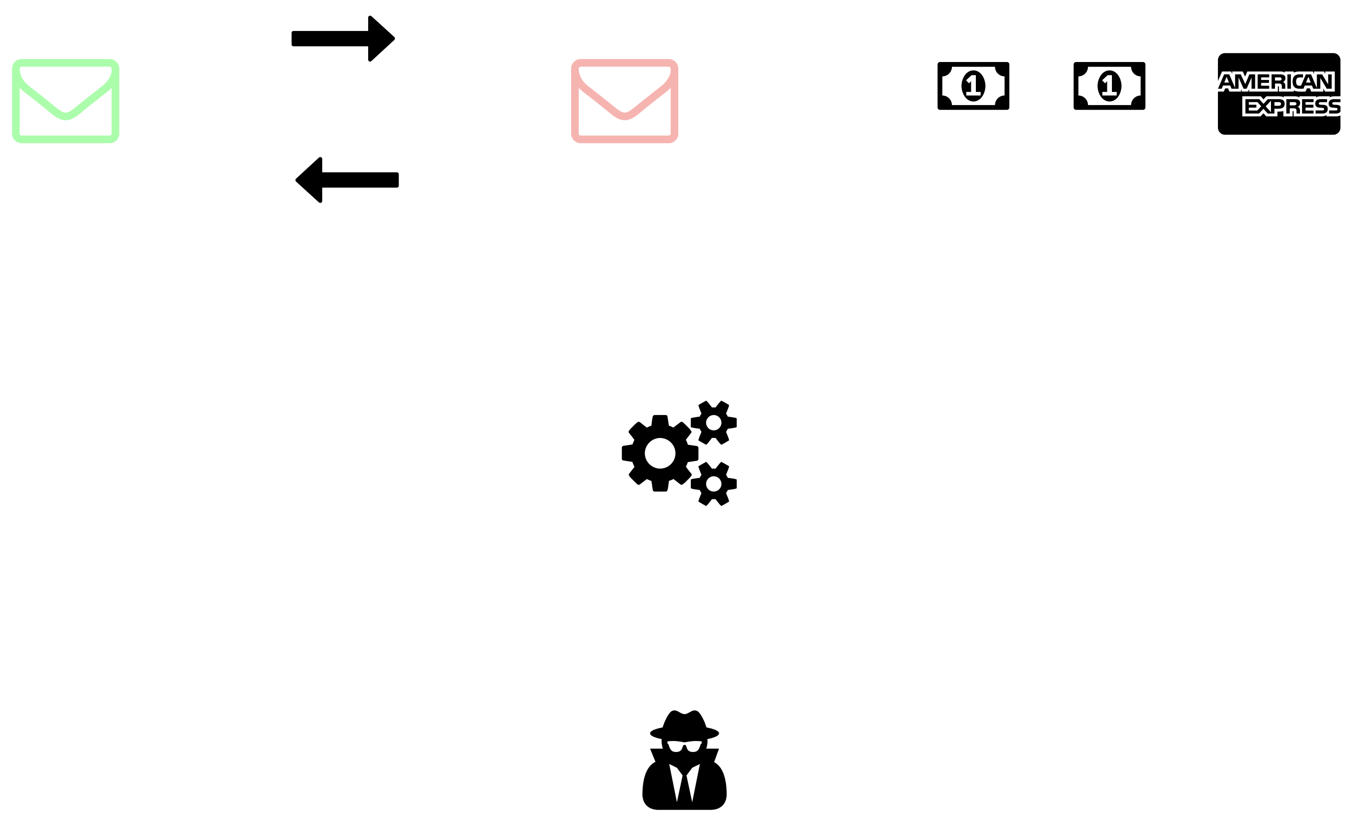
Frequently referenced Washington DC

Frequently referenced Universities

Both Mr.Kaminski and Ms.Mann had raised objections about Enron's dubious financial practices. However, we did not see any conclusive evidence that either of them worked together on exposing this.

Neither Kaminski nor Mann were charged with a crime

Using Machine Learning to Find Financial Criminals



Using Machine Learning to Find Financial Criminals

12 Criminals out of 151 in our dataset

Downloaded Financial data from Kaggle

Clean up financial data

Feature Selection

Financial Data had about 30 features

Salary	Total Stocks	Exercised Stocks
Expense	Bonus	Yearly stock increase

Features we created

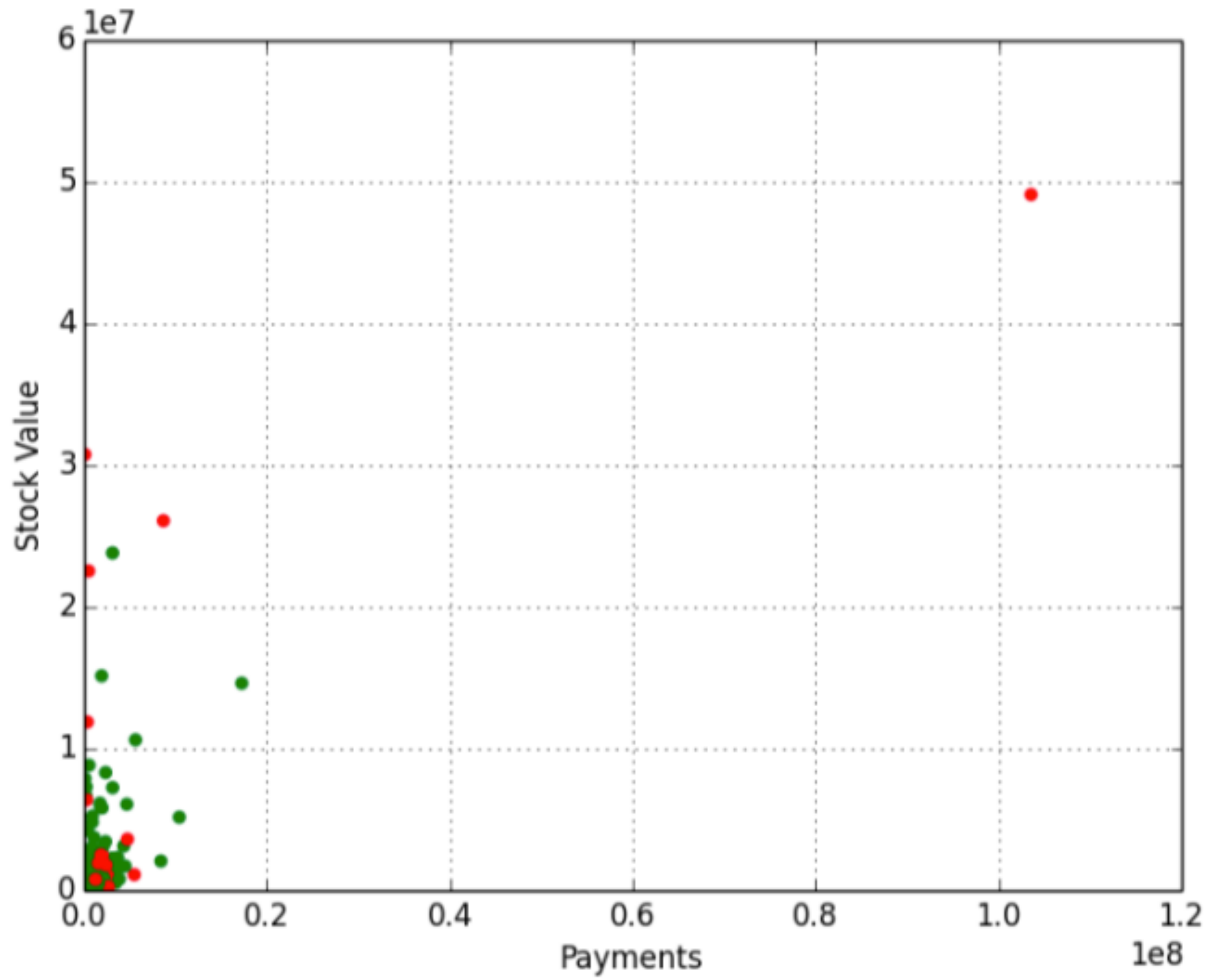
messages_to_con : % of messages sent from user to a person convicted of a crime

messages_from_con : % of messages sent from user to a person convicted of a crime

all_con : % of messages exchanged between user and a person convicted of a crime

Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

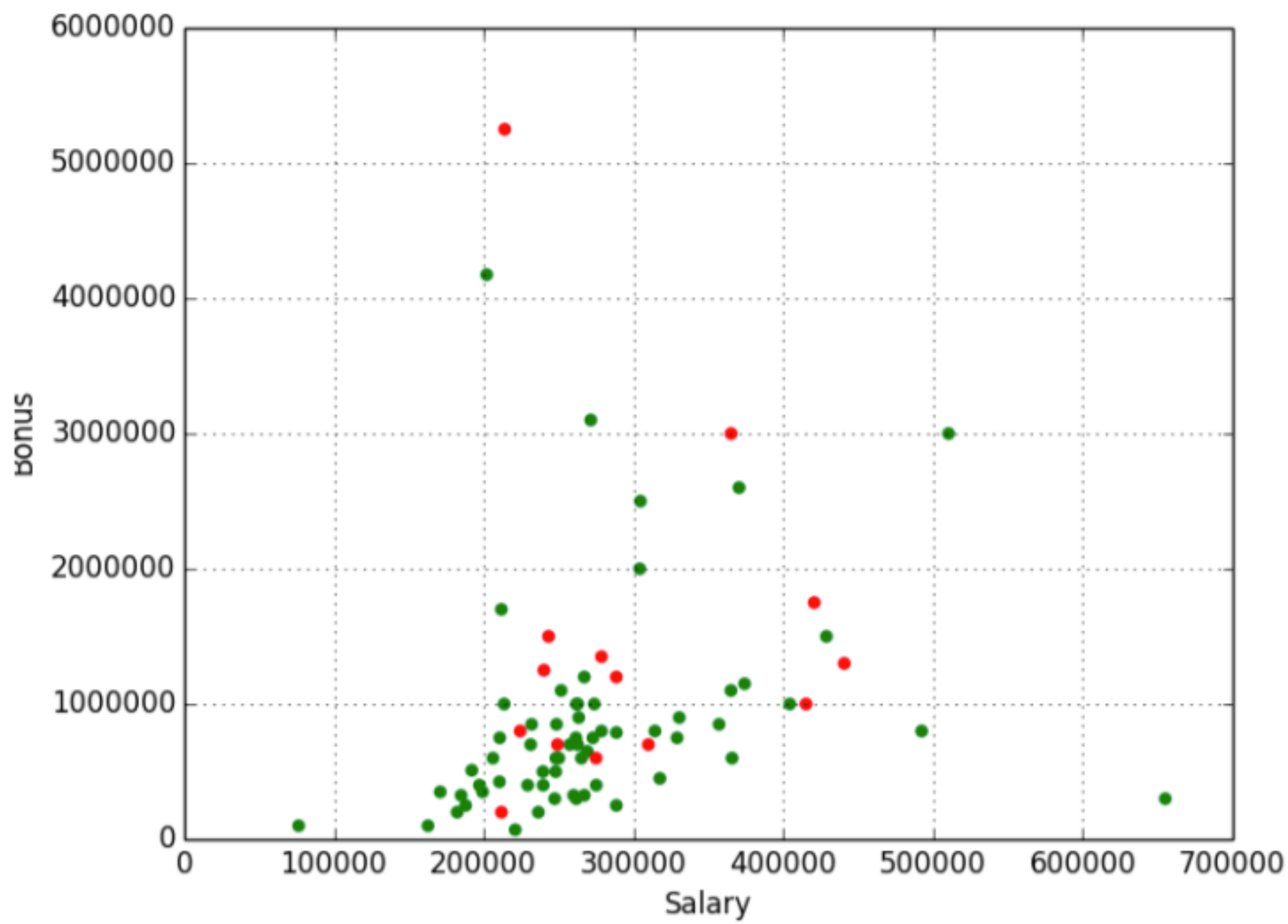


Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

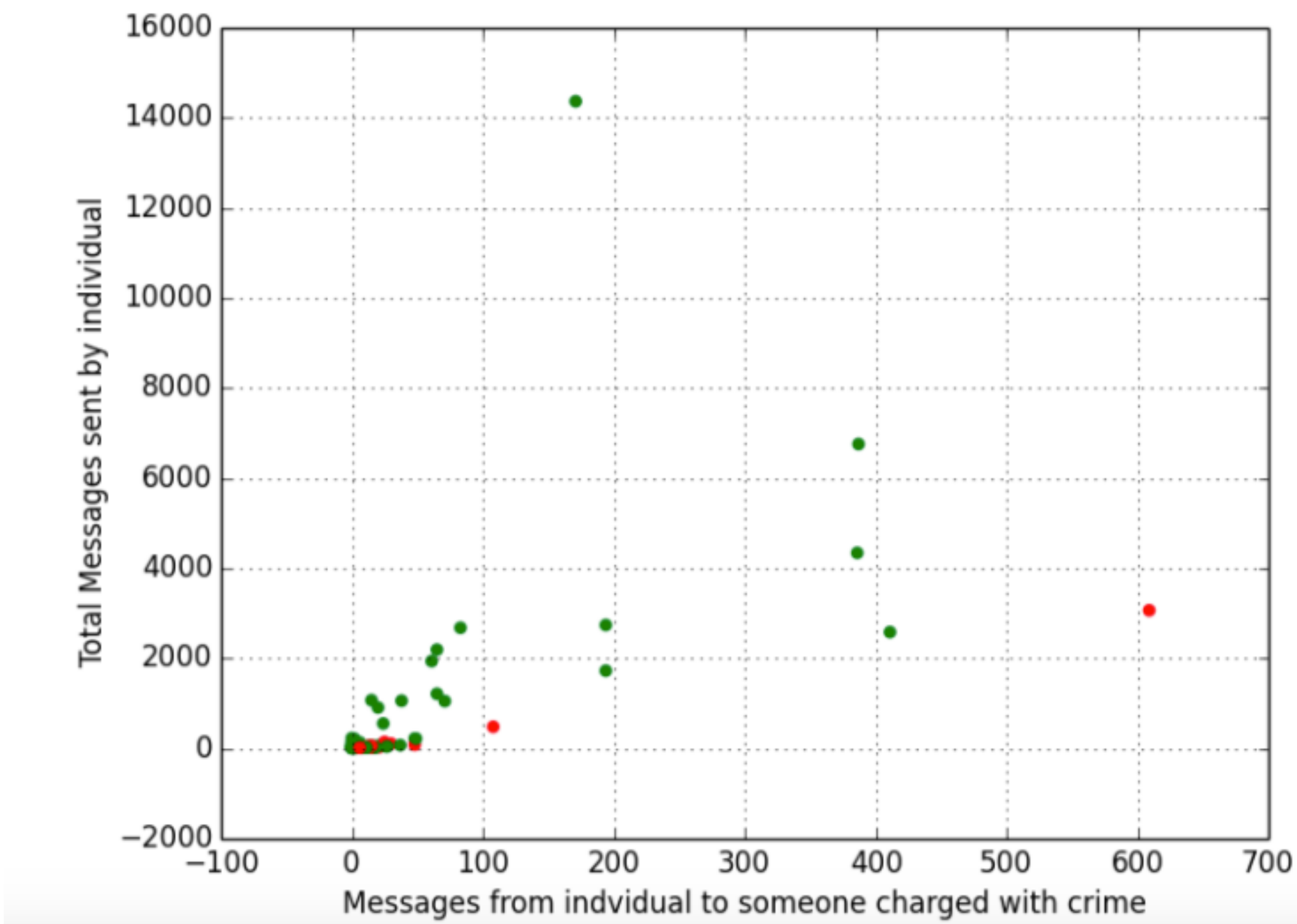


Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset



Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

Exercised Stock Options

75 - 80 %

40 - 45 %

Feature Selection

We looked for features that could clearly separate guilty from innocent people in the dataset

Feature Selection

POI (Class Label)

Salary

Bonus

messages_to_con

messages_from_con

bonus per salary

expense per salary

expenses

exercised stock options

Feature Selection

POI (Class Label)

Bonus

messages_to_con

expenses

exercised stock options

Algorithm Selection

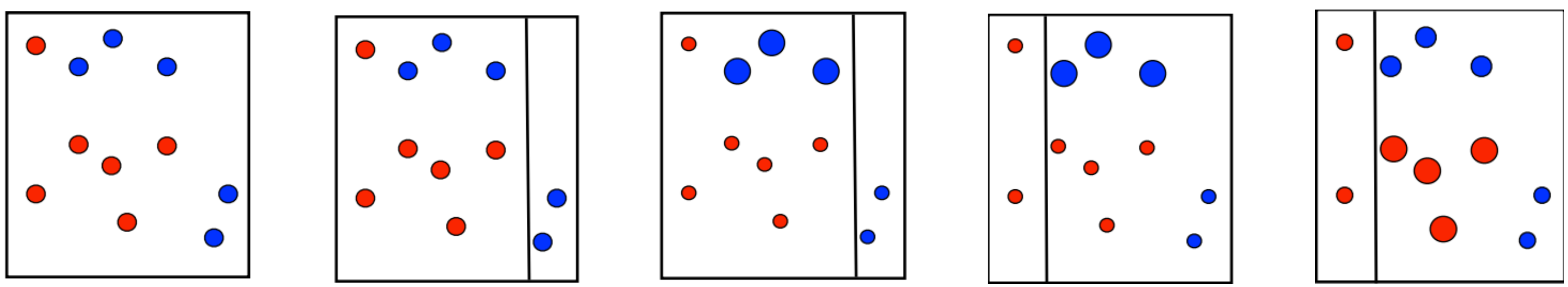
Dataset-size : 151
True positives : 12

From an ML Perspective, building a classifier that does not overfit is incredibly difficult

Ensemble learning methods seemed like the best shot

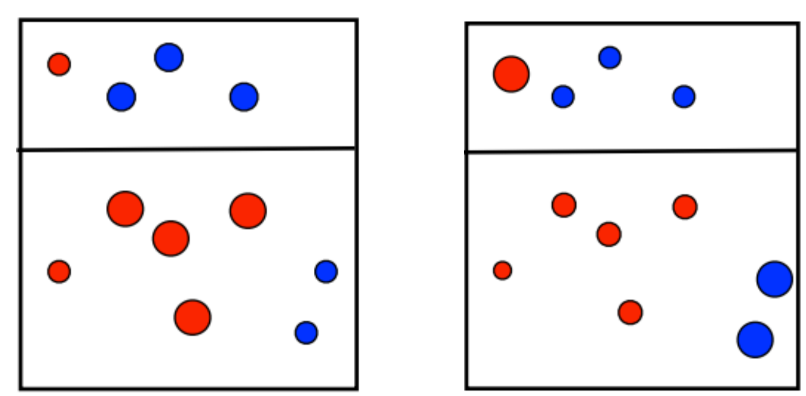
Boosting TL;DR

Use weak learner(s) to create strong learner



Classifier One

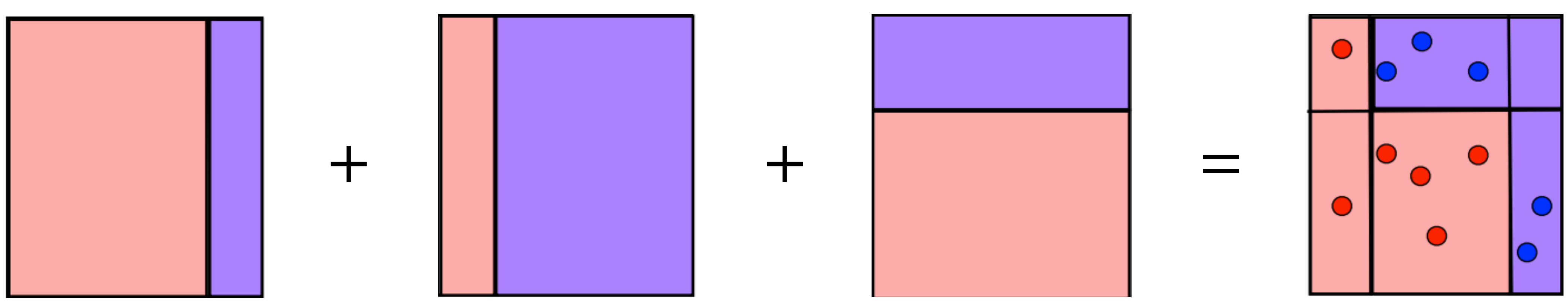
Classifier Two



Classifier Three

Boosting TL;DR

Use weak learner(s) to create strong learner



Bagging TL;DR

Use weak learner(s) to create strong learner

BootStrap **Agg**regation Algorithm

Randomly Sample with replacement N samples
from dataset

Train multiple learners on the data set (may be different)

Average results

Experiments

Used Sci-py to implement both bagging and boosting

Experiments

Used Sci-py to implement both bagging and boosting

Divided dataset into 3 equal parts

Experiments

Used Sci-py to implement both bagging and boosting

Divided dataset into 3 equal parts

Used two to train and one to test

Experiments

Used Sci-py to implement both bagging and boosting

Divided dataset into 3 equal parts

Used two to train and one to test

Representation of innocent and guilty same across parts

Experiments

Used Sci-py to implement both bagging and boosting

Divided dataset into 3 equal parts

Used two to train and one to test

Representation of innocent and guilty same across parts

Ran several thousand iterations

Experiments

Used Sci-py to implement both bagging and boosting

Divided dataset into 3 equal parts

Used two to train and one to test

Representation of innocent and guilty same across parts

Ran several thousand iterations

Varied the number of weak learners

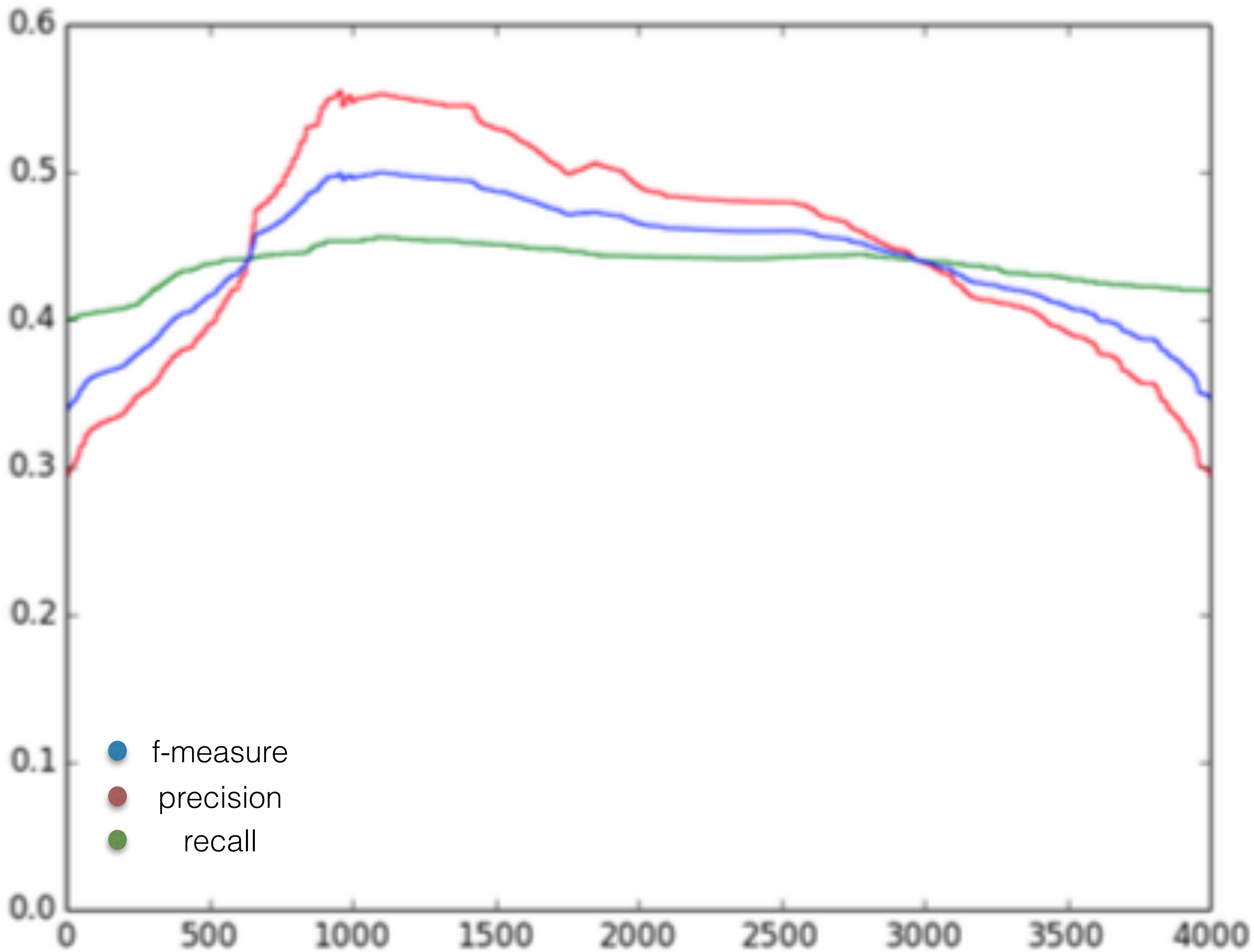
Precision, Recall and F-Measure

Precision : How many of our “true” predictions are right?

Recall : How many of the “true positives” did we find?

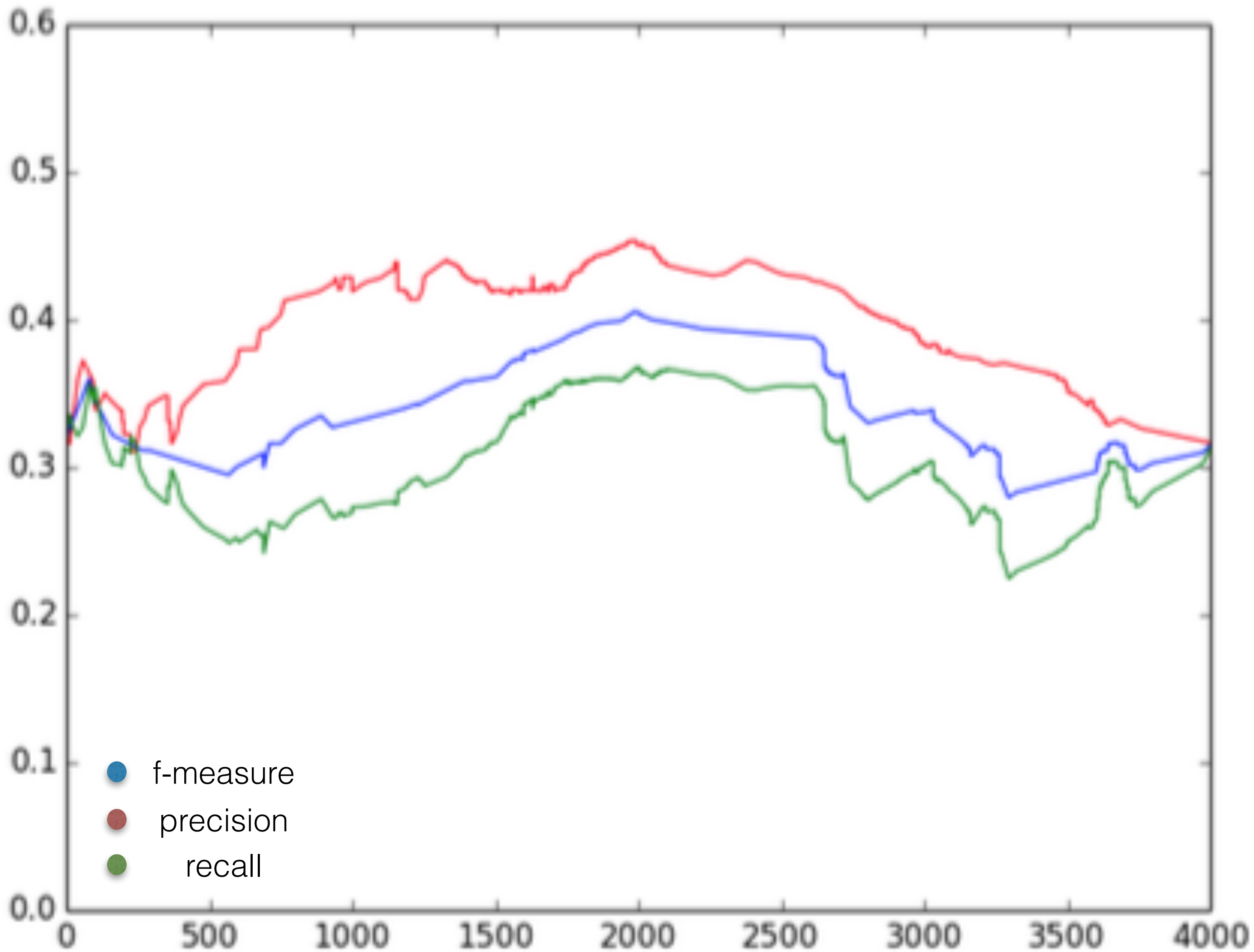
F-Measure : Harmonic Mean of Precision and Recall

Boosting Results



At about 1021 estimators, the f-measure peaks at 0.52 with 56% precision and a 47% recall

Bagging Results



At about 1900 estimators, the f-measure peaks at 0.4 with 45% precision and 36% recall

Bagging vs Boosting

Boosting has clearly outperformed bagging.

This was expected.

Reasons ? *Bagging Boosting and C4.5 - J.R.Quinlan*

In general, even through the f-measures aren't stellar; it is good to know that the system performs **much better** than random guesswork

Sentiment Analysis

We analyzed the sentiments in communication of Enron's CEO(s) (*There were two of them in 2001*) to find whether we could find any indications of whether they knew of the impending collapse.

We parsed through all of CEO's emails in 2001 and developed a classifier that gave us a negative sentiment score for the email. We also used a web based API called *Alchemy* to do the sentiment analysis

Sentiment Analysis